Federal Defenders
OF NEW YORK, INC.

Southern District
52 Duane Street-10th Floor, New York, NY 10007
Tel: (212) 417-8700 Fax: (212) 571-0392

David E. Patton
Executive Director

Southern District of New York
Jennifer L. Brown
Attorney-in-Charge

May 27, 2016

BY ECF
Honorable Valerie E. Caproni
United States District Judge
Southern District of New York
40 Foley Square
New York, New York 10007

Re: United States v. Kevin Johnson
15 Cr. 565 (VEC)

Dear Judge Caproni:

We write to provide the Court with additional support for our application for a subpoena pursuant to Rule 17 of the Federal Rules of Criminal Procedure for the source code underlying FST, as we prepare to file a motion *in limine* pursuant to Federal Rule of Evidence 702 and Daubert. See United States v. Nixon, 418 U.S. 683, 700 (1974).

Mr. Johnson, through the assembled efforts of his counsel, is entitled to access the source code of the Forensic Statistical Tool (FST), the software tool that was used by the New York Office of the Chief Medical Examiner (OCME) to conduct a hypothetical likelihood ratio (LR) experiments, the results of which the government intends to introduce into evidence. Our grounds can be summarized as follows. First, the FST source code is not only the best and most direct, but is truly the only means by which this material evidence can be reviewed and inspected. Indeed, FST is the very source of its manufacture.

Second, and relatedly, the only method available to test the reliability of the LR calculation in this case, with the attendant inquiries of FST's ability to function with specificity, accuracy, and precision, is to perceive its functioning from the back end. This is necessary because of the nature of LR statistics, which postulate the comparison of at least two hypothetical scenarios. No objective means exist to evaluate the reliability of a LR calculation other than to assess the means by which it is calculated: Here, FST itself.

Third, OCME's record of evasive and misleading conduct regarding material issues relating to FST during and after the validation study period for FST substantially undermines the confidence that can be placed in the lab's unsubstantiated representations that the program is functioning without fault. Because of the paucity of information in the public record, published journals, or even retained in the validation study itself, the lab's representations are all the Court has. On this record alone, Your Honor should overrule any objection and allow us to proceed.

Fourth, recently-identified reports within the validation study reveal evidence of FST's imprecision in that it did not reproduce expected results during the validation study. The nature of LR casework makes it impossible to know whether and if so, how, similar imprecision infected LR values reported in this or other cases. The OCME has taken the stance that the discrepancies amount to an error in the data, but that claim is not credible. Additionally, the OCME has argued that the discrepancies only affected the false positive testing in the validation study. Further, the OCME now claims that the software FST used to calculate those LR values is not the same software used to calculate the LR values in casework. These answers only beg more questions and demonstrate our need to inspect the program's reliability.

Lastly, the arguments presented to obstruct our access to the program do not rest on a sound factual basis. The OCME's interest in maintaining perfect secrecy over FST, a program designed with the purpose of creating evidence for use in court, and the very means by which such evidence introduced at Mr. Johnson's a criminal trial was created, must give way to Mr. Johnson's statutory and constitutional rights. Given the efficacy of a protective order, the property and whatever remaining interests can be managed. If it were otherwise, then the requirements of Rule 17 would be read to impose an impossible burden on Mr. Johnson, which is not the standard.

For these reasons, which are discussed in more detail below, and because the affirmative remedy of a protective order available to this Court, the OCME has no cognizable competing interest, our request on Mr. Johnson's behalf for access to the source code for FST should be granted.

I. The Inherent Novelty of this Issue

At this moment in history, there are a relative scarcity of admissibility challenges to probabilistic genotyping software. The lack of cases on point should hardly be taken as evidence of the reliability of FST's technique. The probabilistic assertion is, itself, novel. It has taken time for the court system even to comprehend the nature of the claims made by the manufacturers of probabilistic genotyping systems. It is only in the last four to five years that jurisdictions across the country have begun to see the introduction of LR statistics reported in criminal cases for low-quality DNA associations like that seen in Mr. Johnson's case. The LR statistics offered here may appear weighty, but are in fact an exercise in conjecture, constructed entirely from the OCME's proprietary FST software and data.

As Your Honor is by now aware, the LR paradigm involves highly arcane mathematics and statistics that are confusing even to experienced counsel. This feature of the LR domain is readily admitted by even the strongest advocates for its acceptance. While unfamiliar and complex, these new forensic techniques are still DNA, and thus appear to be associated with a body of largely settled caselaw finding STR DNA and its attendant statistics admissible. The LR is not intuitive. This has produced confusion, such that some early LR cases completely failed to grapple with the unique issues presented by LR statistics, let alone achieve the level of precision undertaken by this Court. See e.g. Commonwealth v. Foley, 38 A.3d 882, 888 (2012)(admitting probabilistic genotyping and mixture-deconvolution expert system TrueAllele pursuant to Frye test and endorsing the trial court's mistaken view that the LR is "a refined application of the product rule.").

In the face of, and in all likelihood equally burdened with these conceptual difficulties, an early groundswell of defense challenges to the implementation of probabilistic genotyping software did not engulf the courts. Rather, in those jurisdictions where pockets of forensic expertise already existed within the defense bar, various challenges have emerged: California, Washington State, the District of Columbia, and also, in response to the rapid implementation of FST by the OCME, here in New York. Notably, at least for the time being, none of these jurisdictions has adopted a Daubert standard.

While Daubert prevails in the federal courts, as compared with state courts, DNA evidence makes comparatively infrequent appearances in the federal system nationwide. Such evidence tends to be introduced in support of classic state offenses, and then is usually reserved for the most serious cases. The number of cases involving DNA in this and the Eastern District, however, is largely attributable to the OCME, as it is one of the comparatively few crime labs in the country able to routinely conduct forensic DNA testing in property, firearms, and other lower-level felony cases.

As the OCME has introduced FST and the LR paradigm into the courts, they have also begun to increasingly use the highly controversial "low-copy" or "LCN" technique. While not dependent upon each other, in many ways, FST and LCN go hand in hand. As LCN produces low-quality samples, rarely of a single source, and rarely representing a complete genetic profile, FST massages sparse data and presents it in such a way as to make it appear more solid than conventional reporting techniques. FST can do this because it compares two conjectural statements. Nothing about FST makes LCN more reliable, simply more palatable.

It appears that many of the infirmities with FST trace to the speed with which it was developed and the shortcuts taken in its validation. The OCME's current efforts to prevent any access to FST find an origin in a pattern of selective disclosures and withholding of key facts about FST and its validation from members of New York State Commission of Forensic Science DNA Subcommittee[1], the official body responsible for the review of FST's validation. These methods speak to an approach to FST's genesis that undermines confidence in the reliability and integrity of the algorithm still in use today.

Genotyping expert systems have come sporadically online for forensic casework over the last four to five years. Their distribution has not been uniform throughout the country. Developed in large part as a response to the intractable difficulties of examiner bias in mixture interpretation[2], many "expert" systems undertake to substitute computerized analytics for the judgment of the lab analyst. The stated object of these systems is to increase efficiency and eliminate biases that have produced horrible injustices, even in the field of forensic DNA.[3] The technology involved in expert systems is as overwhelming as it is bewildering. And for those accused by the computer, it begs the question: who do I cross examine?

---

[1] Hereinafter "the Subcommittee."

[2] See e.g. I.E. Dror & G. Hampikian. Subjectivity and bias in forensic DNA mixture interpretation. 51 Science and Justice, 204-208 (2011) (demonstrating that forensic DNA mixture interpretation is inherently subject to bias and contextual influence).

[3] See http://www.theatlantic.com/magazine/archive/2016/06/a-reasonable-doubt/480747/.

FST is distinguished from other expert systems in that the OCME examiner remains involved as the first step in the interpretive process. FST provides a powerfully-enhanced ability to calculate statistical values associated with a sample. In this, FST inherits all the dangers of bias inherent in classical mixture interpretation. Yet, not grounded in objective standards, it arms the examiner with the ability to calculate exaggerated, misleading, or even flatly erroneous statistics without the examiner ever becoming aware of it. This is because the assumptions and judgments that affect the sensitive LR statistic, judgments that would otherwise be plain and subject to cross examination, are embedded inside the program.

This characteristic of FST may not have been immediately apparent. Early challenges were structured along familiar and not unimportant lines focused on hard laboratory techniques. These processes, while involving matter on the submicroscopic scale, still concerned issues that were inherently "objective." Thus, prior challenges to FST focused on its validation and did not direct resources to the function of its source code. While proper validation certainly is a necessary component of all good scientific technique and engineering, in the case of experimental probabilistic modeling software, in order to test the accuracy and reliability of the process, it is not sufficient to conclude the inquiry after inspecting the validation. It has now become clear that the only means of fully testing the reliability of the process is to inspect the program source code. That is where the work is conducted and the only place it can actually be measured or observed.

Now approaching its fourth year since being validated for coursework, FST code has yet to be reviewed by any court. Because the OCME has not granted access to the source code, none of the results produced by FST have been, or thus far can be, independently verified. Yet, as is discussed below, the record of publication supporting the program is misleading, the empirical basis upon which the program is premised is flawed in its inception, and the statistical underpinnings of the program code are similarly misreported. Recently, evidence has been uncovered in the OCME's own FST validation study demonstrating problems with FST's ability to reproduce expected results. This is the paramount expectation for a process demanded to accurately handle large and complex equations. Yet, even before this Court, the OCME has encouraged flatly misleading representations that portray FST as earning sterling validation credentials and built from bullet-proof source code. These tactics may have succeeded for the OCME in the past, but they should not suffice here, as they do not have the support of the facts.

II. General Problems with FST Underscoring the Need for Access to the Source Code

Instructed by one published article[4], it remains unclear how FST is achieving its results. In broad strokes and perhaps even as expressed through abstract algebraic symbols, the theory behind the program is clear enough, and undoubtedly novel, although of questionable merit applied here. Applying fixed estimates of drop-out and drop-in, derived from artificial simulations conducted under pristine conditions to complex forensic mixtures, the process depends for its validity upon an the analyst's accurate assessment of DNA sample template amount, maximum estimated number of contributors to the mixture (which cannot under any circumstances exceed three), and an estimation of the ratio of each contributor to

[4] See Adele Mitchell, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. FORENSIC SCI. INT'L: GENETICS 6. August 9, 2012 [hereinafter "Mitchell 2"], Attached at Exhibit A.

the mixed sample.[5] The FST process thus begins with the examiner's interpretation of the mixed sample, including the inescapable effect of examiner bias, but is then rigidly lashed to the preset variables. Never incorporating information no matter how plainly observable in the data, FST instead enforces a willful analytical blindness to the case evidence in favor of the preset rates of drop-out, drop-in, and other variables embedded in the source code. FST is unique in this way. No other program biases LR calculations against objective, specific, and evidence-based judgments towards generalized, "empirical", tabulated data in the way FST does[6] This scrupulous avoidance of evidence-based reference to the case data would seem an obvious forensic deficit, but in the hands of FST's advocates, it is promoted as a strength. To FST's authors, the drop-out rates are "empirically derived" from the validation study.[7] This, of course, should appropriately place a keen focus on the accuracy and validity of drop-out study. It is doubtful that it can withstand sustained scrutiny.[8]

III. FST Source Code is the Primary Source of the Evidence Against Mr. Johnson

We have made some effort to show that even in the field of forensic probabilistic genotyping software, the OCME's "black box" practices are outside the scope of good science, particularly for a public crime lab, which depends on open inquiry to develop. The OCME's information embargo is certainly at odds with fundamental fairness, due process, and confrontation rights, and now threatens compulsory process. This secrecy, supposedly to shield property rights not truly threatened by Mr. Johnson, only reveals the nature of the difficulty in which the defense finds itself.

We agree that FST is a singular technical process that cannot be replicated by other means. It is precisely this uniqueness that creates the need for access to FST's source code. If the actual true algorithm were otherwise available, if, for instance, the LR was simply a matter of summing known variables on an Excel spreadsheet, then the nature of this dispute would be vastly different. We would be entitled to be informed of the values for each variable but could let Excel perform the task of doing the math, expecting Microsoft to have reliably developed its source code.[9]

Here, though, the OCME has not just chosen the variables in FST but obscured each value, publishing vague and often contradictory claims about their purpose and operation within the algorithm. The lab has engineered the software, and secrets the true functioning of the complex mathematics that produce a report intended for admission into evidence – not general research or an investigative lead – yet screens from all review the experiment by

---

[5] Id.
[6] Mitchell 2 at 759.
[7] Id. at 750.
[8] See infra at p. 9-13.
[9] Of course, even Microsoft, a multibillion dollar company dedicated to the production of commercial software, encounters unexpected problems with precision in the source code in some of its products, particularly when it comes to complex calculations, as they have with the floating point calculator function in the Excel spreadsheet. See https://support.microsoft.com/en-us/kb/78113. Attached at Exhibit B.

which that evidence was created. It is not as if reagents have been consumed. The source code exists and remains in continuous use.[10]

IV. Access to the Source Code is the Only Means for Testing the Reliability of FST

A. Lack of Technical Review

It is not possible to reproduce the results of the FST experiment or further test the reliability of the true method the OCME utilized to produce the LR in this case without access to the source code.[11] Yet, the OCME has never allowed independent review to ensure that its true functions corresponds with the processes described in the limited published literature, or otherwise satisfactorily established the reliability of the program. The validation study is not sufficient to demonstrate functional accuracy. Reams of reports of LR values that are "generally concordant" with broad categorical "manual interpretations"[12] do not support the proposition that those LR values are either accurate or reproducible. Nathanial Adams, a Systems Engineer retained by the defense explains:

> ...FST's validation studies, comparing various mixed samples constructed from known contributors combined at certain precise proportions, and observed at a set number of precise template amounts ... were designed to test for the occurrence of stochastic effects on differing qualities and quantities of mixed DNA template. These results do not clarify the specificity of the LR producing algorithm employed by FST or the quality or accuracy of the software development processes utilized to translate this algorithm from scientific concepts into source code.[13]

Beyond conceptual issues with FST, beyond the question of how those issues affect the LR calculation within the source code, there is the question of whether the development of the program accurately conveyed FST from concept to execution.

As with all probabilistic genotyping software, the modeling assumptions upon which FST is predicated compare hypothetical scenarios and thus cannot be verified against

---

[10] The government's attempt, presented without standing on the Rule 17 issue, to distinguish Mr. Johnson's need to review the source code from copyright and patent cases is unavailing. The property rights and interests in those cases is certainly of no greater weight than the constitutional rights at stake here. Further, those cases only demonstrate the ability even of competitors in copyright and patent litigation to submit such material under protective orders. The mystifying LR statistics offered here cannot be reconstructed outside of FST, and OCME's claim to trade secrets to obstructs the defense from conducting what would otherwise be standard review and testing of the evidence. FST may be novel, but it has not ushered in a new era of diluted due process.

[11] Accord Declaration of Dr. Ranajit Chakraborty in support of Motion to Suppress, at ¶ 30, United States v. Rashawn Jermaine Smalls, No. 14-414 (BMC) (E.D.N.Y. 2015), ECF No. 29-2 (hereinafter "Chakraborty Decl.") infra p.7 (We have retained Dr. Chakraborty, but the Federal Defenders are not associated with the Smalls case or defense. However, as his affidavit in that case is instructive of many of his insights and informed opinions on FST and its source code, we cite to his affidavit in the record of that case.)

[12] Id. at 760.

[13] Declaration of Nathanial Adams, Attached at Exhibit C, at p. 2. ("Adams Decl.")

objective standards.[14] To the extent that these assumptions are embedded within the source code, a review conducted by a competent Systems Engineer can test FST's back end for its adherence to its stated aims. This is the only reliable and conclusive method to engage the larger inquiries of accuracy and validity of the LR calculations in this case. Indeed, according to defense expert Dr. Ranajit Chakraborty, a man of considerable experience in the field, including many years of close association with the Federal Bureau of Investigation's Crime Lab, any other means to attempt such an inquiry would be "virtually impossible."[15]

B. Lack of Sufficient Peer Review

There are only two published articles on the subject of FST, one of which contains not more than a page of text and is largely reproduced in the subsequent publication.[16] The relative paucity of publications regarding FST in the scientific literature as compared to other probabilistic genotyping software tools[17] contributes to the impenetrability of the program to meaningful review.

---

[14] Id. at p.1; accord C. D. Steele & D. J. Balding. Statistical evaluation of forensic DNA profile evidence, ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION 1, 2014, at 361-384.

[15] Chakraborty Decl. at ¶ 30.

[16] See Mitchell 2, see also Adele Mitchell, et.al, Likelihood ratio statistics for DNA mixtures allowing for drop-out and drop-in. FORENSIC SCI. INT'L: GENETICS 3. August 31, 2011 (two pages, totaling eight paragraphs, only two of which included information not later provided in Mitchell 2, and these were case summaries) (hereinafter "Mitchell 1"), Attached at Exhibit G..

[17] Compare Mitchell 1, Mitchell 2 with M.W. Perlin. Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information. 6 J. PATH. INFORMAT. 1:59 (2015); S.A. Greenspoon, et al. Establishing the limits of TrueAllele Casework: a validation study. 60 J. FOREN. SCI. 5:1263-1276 (2015); Perlin, M.W., et al. TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors. 60 J. FOREN. SCI. 4:857-868 (2015); M.W. Perlin, et al.TrueAllele Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. 9 PLOS ONE 3:e92837 (2014); M.W. Perlin, et al. New York State TrueAllele Casework validation study. 58 J. FOREN. SCI. 6:1458-1466 (2013); J. Ballantyne, et al. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. 53 SCI. & JUSTICE 2:103-114 (2013); M.W. Perlin. When good DNA goes bad. J. OF FOREN. RESEARCH. S11:003 (2013); M.W. Perlin, et al. Validating TrueAllele DNA mixture interpretation. 56 J. FOREN. SCI. 6:1430-47 (2011); M.W. Perlin & A. Sinelnikov. An information gap in DNA evidence interpretation. 4 PLoS ONE 12:e8327 (2009); M.W. Perlin, et al. Match likelihood ratio for uncertain genotypes. 8 LAW, PROBABILITY AND RISK 3:289-302 (2009); S.Y. Hill, et al. A genome-wide search for alcoholism susceptibility genes. AM. J. OF MED. GENETICS PART B (NEUROPSYCHIATRIC GENETICS):102-113 (2004); K. Kadash, et al. Validation study of the TrueAllele automated data review system. 49 J. FOREN. SCI. 4:1-8 (2004); M.W. Perlin & B. Szabady. Determining sequence length or content in zero, one, and two dimensions. 19 HUMAN MUTATION 4 (2002); M.W. Perlin & B. Szabady. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. 46 J. FOREN. SCI. 6:1372-77 (2001); B. Pálsson, et al. Using quality measures to facilitate allele calling in high-throughput genotyping. 9 GENOME RESEARCH 10:1002-1012 (1999); G. Lancia & M. Perlin. Genotyping of pooled microsatellite markers by combinatorial optimization techniques. 88 DISCRETE APPLIED MATH 1-3:291-314 (1998); M.W. Perlin, et al. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. 57 AM. J. OF HUMAN GENETICS 5:1199-1210 (1995); M.W. Perlin, et al. Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. 55 AM. J. OF HUMAN GENETICS 4:777-787 (1994).

Dr. Chakraborty, a widely respected expert in the field of Molecular Genetics[18], in a 2014 Eastern District case not involving the Federal Defenders was asked to review LR evidence produced by FST. In that case he made the stakes of the FST source code issue perfectly clear: "Without knowing the source-code, it is virtually impossible to test the reliability of the LR calculations performed in [the] case. OMCE's [sic] publications to date are insufficient to assess reliability."[19] For a man who had sat on the DNA Subcommittee during the entirety of FST's putatively rigorous validation process, who had built a career establishing the scientific and technical foundations of forensic DNA typing, and who had testified on behalf of the government in nearly all of his courtroom experience, this unequivocal statement is notable. Having been involved in the review of FST's validation, Dr. Chakraborty understands the program and its theoretical underpinnings very well. His assessment that it cannot be evaluated short of a review of the source code should be attributed significant weight.

He also has grounds for concern: Dr. Chakraborty voiced concerns during FST's review process, particularly relating to the drop-out study. Significant time has passed since these exchanges. Yet, the OCME still has not addressed those issues. "Despite OCME's own admissions of weakness in the FST program, to date, OCME has not published any additional studies documenting how case-specific variables may affect the LR. Nor has [the] OCME made any documented adjustments to the FST program."[20]

Moreover, to date, the OCME still refuses to submit FST to any kind of meaningful independent technical or scientific testing or review. From appearances, the lab has chosen to rest upon the validation study. Thus, Nathanal Adams: "I have never seen formal descriptions of software quality assurance or software testing plans for FST, such as are common in the field of software engineering, and which would increase clarity about intended vs. actual performance of the FST software system."[21]

If this were mere reticence, it would be less troubling. But the limited literature that is published on FST is stock with misleading ambiguities that invite the Court – and any audience – to conclude that FST has been designed with a care that it has not, and that the program has been subject to rigorous verification when it never has.

C. FST's Troubled History Casts Doubt on Unsubstantiated Claims as to Its Reliability

In all likelihood, FST does not have a bright future. The scientific basis for adhering to a set of "empirically derived" variables in calculating a single LR value for introduction in a criminal case has not gained wide acceptance outside the New York State, where the OCME has a something of a lock on the market. Information has also become widely known that

---

[18] Dr. Chakraborty is perhaps most notable for having assisted in the development of the 13 STR forensic genetic markers upon which the entire basis of modern forensic DNA typing is founded. Dr. Chakraborty also served as a member of the New York State DNA Subcommittee from its formation in 1995 until 2011. In his capacity as an expert, Dr. Chakraborty has testified in over 200 cases, approximately 95% of the time on behalf of the prosecution. Chakraborty Decl. at ¶ 8.

[19] Id. at ¶ 30.

[20] Id., ¶ 33.

[21] Adams Decl. p.2-3.

the OCME has purchased a large number of site licenses for the "competitor" probabilistic genotyping software product STRmix, and is currently validating it for use. One of FST's chief authors, Theresa Caragine resigned from the office amidst accusations of not following protocol regarding the reporting of casework of a subordinate, and Adele Mitchell now works in private industry. That the program may yet be scuttled does not excuse its troubled history or shoehorn its continued use without necessary access to its true function.

(i) Misleading Characterizations in FST Publications

In 2012, when the Mitchell 2 article was published, very few in the court system had encountered FST or the LR paradigm for expressing "weight of evidence" statistics despite evidence of an exclusion. FST's authors crafted the article with a mix of bewilderingly detailed processes and formulae, impressively large numbers, and calming language that appealed to "conservatism." To this mix was added a series of opaque statements covering what would turn out to be the most controversial issues involving FST.

For the layperson in the court system, the charts may not help acquire an informed understanding of FST's true function[22], as they provide more detail than description. Similarly, the aggregate numbers do not assist the reader in coming to an understanding of obviously pertinent questions, like the size of relevant sample groups.[23] Worse, the aggregate reporting of numbers tends to obscure the fact that the "sum totals" are not the relevant numbers.[24] At times it appears that numbers are included in the article for rhetorical effect. In one confounding passage, under "2.2.1 DNA Sample Collection" at page 752, the authors list the numbers of samples collected of five types of mixed samples comprising the "test set of mock evidence samples." Adding these five groups together produces a total of 391 mixed samples. Yet, within the same paragraph, the authors clam that "350 mixtures...were included in the validation." The missing 41 samples are not further explained.

The language of conservatism is similarly confounding. As pointed up by Nathanial Adams, the repeated usage of the word "conservative" in the publication does not equate with a clear definition of the term:

> I have seen the claim by Mitchel, [sic] et al. that FST's LR calculations were "more conservative than those performed by hand". The term "conservative" is used five times in this article without a definition. Different definitions of "conservative" can lead to drastically different interpretations of this statement. The impact of various definitions of conservatism on reported likelihood ratios in validation studies and casework samples could be more thoroughly investigated with access to the FST software system and source code. For example, the effects of adjusting the described "conservative" statistical models could be measured regarding reported likelihood ratios and rates of false exclusions and inclusions.[25]

---

[22] See. e.g. Table 3 "Factors used in denominator" p. 752.
[23] Id. at 750.
[24] Infra § ii, p. 10.
[25] Adams Decl. at p. 3.

In truth, Mr. Adams is generous to the authors. While the term "conservative" does appear five times in the article, the word is used at directly cross purposes. In ordinary scientific usage, the word "conservative" calls judiciousness and forbearance to mind, and here would lead the reader to assume that the authors were employing methods chosen to safeguard against LR values that would overstate the weight of an association. This view is reinforced by the first appearance of the term on page 752 of the article, where, "[i]n order to be conservative, FST uses the drop-out rate minus one standard deviation[.]"(emphasis added). The authors' approach to the standard deviation differs from the purportedly conventional method of subtracting two standard deviations, a step which would result in much lower drop-out rates, and thus produce much higher LR values. This would prejudice the defendant by attaching greater statistical weight to an association. From such appearances, the authors were taking a conventionally conservative approach to the design of FST.[26]

Yet, by page 758, the meaning of "conservative" had been inverted, now describing the OCME's choice to use only its most sensitive instruments to conduct the testing in the drop-out study. The choice to use these more sensitive instruments to set the rates would produce a clear under-estimation of drop-out even compared with testing performed on OCME's average-sensitivity instruments, and a more pronounced effect for that performed on their least sensitive instruments. An underestimation of drop-out increases the likelihood of a false positive, over-values the weight of an association, and thereby favors the prosecution. In choosing to conduct the drop-out study on the most sensitive instruments, the authors construed the term "conservative" in the opposite manner than they had just pages earlier. It was in fact a highly anti-conservative choice to go with the most sensitive instruments. Nevertheless, the authors labeled their choice "a conservative approach."[27]

There can be little doubt that the authors understood choices they made in the study, or in drafting the article. Even a lay reader may understand the concept of a standard deviation, but to parse the significance of utilizing highly sensitive testing equipment on the ultimate LR calculation would take years of experience or expertise. At the time of the article's publication little of either existed in the court system.

(ii) *Misleading Size of Samples in Drop-Out Study Undermines Function*

Presenting the drop-out study in the article, the authors report the constituent samples as aggregate amounts. (i.e. "In total, more than 2000 amplifications";[28] "In total, 350 mixtures and 104 touched items";[29] "In total, more than 557,00 test runs of the program were performed[.]")[30] Though perhaps impressive on their face, these grand totals are not the relevant numbers.

---

[26] We do not concede that anything about the drop-out study was in fact "conservative." Indeed, the employment of one, rather than two standard deviations appears to have more to do with inadequate sample size, but has not been adequately explored for the purposes of the Rule 17 motion.
[27] Mitchell 2 at 758.
[28] Id. at 751.
[29] Id. at 752. When presenting the number of different categories of mixed samples, the sum total the authors present exceeds 350, which is not possible to reconcile from the text of the article. Id.
[30] Id. at 750.

Considering the design of the drop-out study, which was predicated on the independence of drop-out rates for each comparison group, the overall number of amplifications is insignificant. Rather, it is the size of the comparison group that matters for statistical, population, and other research purposes. Starting with the 2,000 amplifications, that group was divided into at least 64 comparison groups. If the parameters identified in Mitchell 2 are to be taken as the guide[31], testing was broken down between single-source and mixed samples. Single-source samples were then broken out into 8 testing groups by template size. Mixed samples were broken out into 2 and 3 contributor groups, each of which was further broken down once again into 2 ratio groups. Each of these resulting groups was then broken out into 6 testing group by template size. The entire array of single and mixed samples was broken out again by number of PCR cycles, one HT-DNA group subjected to 28 cycles, and the other LT-DNA group subjected to 31 cycles. The 2,000 amplifications in the study, divided by the 64 subgroups produces approximately 31 amplifications per group, a far smaller number per comparison set. This is not reported in the literature.

(iii) Opacity Regarding the Central Issue of Degraded Samples

Further, while the published literature makes repeated reference to the incorporation of degraded samples in the article is opaque as to whether the data from those degraded samples would be used to the drop-out rate estimations employed in FST: Application of FST to degraded casework samples was and remains a common issue. Yet, the authors do not reach a conclusion.

> Ultimately, it was determined that, in general, use of the degradation module as programmed resulted in LRs closer to 1.0 for both true contributors and non-contributors. That is, this approach did not increase the overall separation between true contributors and non-contributors (data not shown).[32]

Referencing the validation study, however, in a document finalized with the authors more than a year prior to the date Mitchell 2 was submitted for publication, reveals that these results were not in fact intended to be used for casework:

> NOTE: A set of purposefully degraded samples was used to establish degradation specific drop-out rates. However, the utilization of these rates proportionally increased the LR value for both the true and the non-contributors, thus not aiding in distinguishing between these groups. Therefore it is not planned to apply these drop-out values for casework.[33]

---

[31] There is evidence in the validation studies to suggest that the number of amplifications reported in Mitchell 2 was divided among a greater number of subgroups. For example, in Volume 10, the Identifiler 28 cycle runs were performed on a 3-contributor mixture set comprised of 3 proportions, not 2, as described in Mitchell 2. In Volume 8, Identifiler 28 cycle runs were performed on a 2-contributor mixture set comprised of 7 proportions, not 2, as claimed. In both Volumes, the additional proportions are explained as being extant for other purposes in FST testing. Nevertheless, the key question is whether their amplification was counted against the total number of 2,000 amplifications claimed in Mitchell 2 at 751.

[32] Id. at 759.

[33] Executive Summary, FST Validation Study, ("Ex. Summ."), at. p.10, attached at Exhibit D.

This clear conclusion is not reported in the literature. Yet, references to testing sets of degraded samples of various amounts (i.e. 93[34] and 97[35]) appear throughout the article. It remains unclear how large the rejected "degradation module" was, and whether the samples were included in the reported group of 2,000 amplifications. Fundamentally, that "a set" sizeable enough to constitute its own study was jettisoned for the purposes stated provokes further questions, here the most relevant being the level of attention paid by FST source code architects to basic statistical principles.

This exercise highlights concerns about the carelessness of the authors of FST's source code. It is difficult to understand why the degradation module is not employed by the OCME in casework, given the authors' concessions about the reality that degraded DNA is encountered in everyday forensic practice[36], or how, in light of this understanding, a drop-out rate incorporated into FST predicated on pristine buccal swabs tested on the most sensitive instruments could be considered "conservative". The article is not instructive where it counts, presenting trivial numbers like 391 mixed samples obtained while only 350 were used, while not reporting consequential data like the number of comparators in each relevant sample group, or the size of the degradation module, or any meaningful information regarding the actual rates of drop-out that were purportedly obtained from the study. Taken together, the authors' choices undermine confidence in their representations about the source code, illustrate a rush to put FST on-line for casework, and illustrate a practice of artful non-disclosure regarding damaging information.

(iv) Samples in the Validation Study are Artificially Hybridized

Perhaps the most glaring omission in the article also does not appear in the validation study, and was not disclosed during the validation process before the Subcommittee itself, and relates to the very samples used to conduct the study. Subsequent to its presentation to Subcommittee, it came to light that the study authors had not truly amplified 15-locus profiles for each of the samples in the study. Rather, they had amplified a number of profiles used in the study only to 13 loci, a standard previously in wide use. To incorporate the additional loci, D2S1338 and D19S433, into the study, rather than re-amplify the samples, the authors simply "simulated genotypes for these two additional loci and included them in the multi-locus DNA profile date in their validation studies."[37]

Thus, an unknown number of profiles used in the study do not represent actual people. This obviously impacts the validity of any inferences that could ever be drawn from the study as to population statistics. According to Dr. Chakraborty, this hybridization is "statistically unacceptable because it produces pseudo-independence of genotypes of some loci in [the] data."[38]

The authors matched the boldness of this choice with the subsequent, greater omission: They did not disclose this fact to the Subcommittee. "[The] OCME did not reveal

---

[34] Ex. Summ., p.10.
[35] Mitchell 2 at 752.
[36] Id. at 759.
[37] Chakraborty Decl. at ¶41.
[38] Id.

this approach during its presentations to the DNA Subcommittee. ... If this fact was known during my participation on the DNA Subcommittee, my decision could have been drastically affected."

V. Further Misrepresentation about Checking the Math

Sadly, the misrepresentation of FST's validation made their way into the record of this case, having been presented directly to this Court at the May 9, 2016 oral argument. Presumably having first consulted with the OCME, the Government asserted that the OCME "checked all the math by hand" of FST's LR calculations during the program's validation.[39] Further claims were made that "what is happening here" is that the defense experts don't want to do the math by hand.

Provocative. Not true. Neither the literature nor the validation study support government counsel's expansive claim about the level of mathematical hand-checking accomplished by the OCME. The record doesn't support counsel's claim because it did not happen that way. The claim would in fact be impossible. Appearing repeatedly in the scholarship discussing probabilistic genotyping systems is the recognition that LR calculations for complex mixtures are incredibly onerous and in most instances could not possibly be done by hand. The Scientific Working Group on DNA Analysis Methods (SWGDAM) Guidelines for the Validation of Probabilistic Genotyping Systems, approved as of June 15, 2015, lend authority to this well-understood casework reality. In Guideline

---

[39] Transcript of Oral Argument, United States v. Kevin Johnson, May 9, 2016, 6:17-20

("But they checked all the math by hand with the very same formula that defense counsel has").

The Court: So are you saying that part of what these validation studies say is that they did, on a calculator or whatever, they ran samples through FST and they did the same calculations by hand?

Mr. Swergold: Yes.

The Court: To verify?

Mr. Swergold: Yes.

The Court: And they came out the same?

Mr. Swergold: Yes. That's part of the validation study. They did manual calculations by hand to check the work that FST was doing.

The Court: So for all of the many calculations that the computer has to do, they checked them manually?

Mr. Swergold: I don't know whether they did for every single test that was run, but they did it as part of their validation study.

Id. at 42:21-43:11

3.2.6.1, discussing internal laboratory developmental validation procedures, recommends that efforts should be made to compare the results of probabilistic genotyping software to manual results or alternative software packages to aid in assessing the accuracy of the system. The authors go on to note "[c]alculation of some profiles (e.g., complex mixtures), however, may not be replicable outside of the probabilistic genotyping system."[40]

It is simply not true that the OCME performed the math by hand for anything close to the 557,000 LR calculations executed in the course of the FST validation study. If the entire staff of the OCME's office had undertaken this project in 2011, they would not yet have completed the job. Returning to the literature, as with other material issues in the validation study, the Mitchell 2 article is opaque and potentially misleading here. The authors employ two closely-related terms, "hand calculations" and "manual verification", to describe materially different steps in the validation processes.

One term, "hand calculations", refers to the kind of mathematical work-checking that was represented as being extensive: "The final program was tested by performing hand calculations to verify the expected result based on the algorithms explained above and the user set sample characteristics."[41] Standing alone, the reader could conclude that checking FST's work was a significant portion of the validation. That conclusion would be inaccurate.

The Executive Summary describes this step as "manual calculations": "Program performance was verified by comparison of manual calculations to program output [such that the intermediate steps were visible] for twenty-four examples."[42]

Twenty-four incomplete calculations is not anything like what has been urged on the Court. Expressed as a percentage of the reported number of LR values calculated in the study, it is a miniscule 4.309e-5. This purported "testing" can only be understood as not supporting the government's claim.

In addition to the obvious point that twenty four calculations amounts to hardly more than a spot check, the key information here is the bracketed conditional statement: "such that the intermediate steps were visible." All that can accurately be deduced from this statement is that the 24 LR calculations were not complete models of the FST algorithm, but akin to check-ups of pieces of the math. How many, and how extensive these calculations were is not reported anywhere.

The other "manual interpretations" referenced on page 760 were not mathematics at all. Rather, they were qualitative interpretations, not calculations. The process was akin to a qualitative assessment of the electropherograms ("epg's") produced in a mixed samples. Reviewing the epg, the analyst assessed whether the profiles of each contributors could be deduced. The analyst recorded their assessment for each contributor as one of 5 pre-defined qualitative categories: "major components", "included", "cannot be excluded", "no conclusions", or "excluded".[43]

---

[40] http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf (emphasis added)(last visited May 27, 2016) at p. 7.
[41] Mitchell 2, at p. 752.
[42] Ex. Summ. p.3.
[43] Ex. Summ. p. 11.

After the LR's were run for all samples, the qualitative manual assessments and the LR results were cross-tabulated in tables 3A and 3B of the study. As can be seen in the pages of Table 3A[44], in the "manual call" columns, one of the 5 pre-defined qualitative categories are assigned in each row – yet they have no precise scientific correlation to the LR for the same sample in the column next to it. According to Mitchell, the results were "generally concordant."[45] This is not a statement of mathematical equivalency, nor could it be, as the qualitative assessment is not numerical. Likewise, according to the Executive Summary: "Overall, LR values were consistent with manual conclusions."[46]

The qualitative assessments in tables 3A and 3B, combined with the 24 evidently incomplete calculations of LR components which represent a miniscule fraction of the number of LR values calculated in the study, together comprise the entire set of the "testing" comparisons conducted during FST's validation. The published literature is opaque as to both the quality and the quantity of these comparisons, but the validation study record reveals that the OCME simply did not engage in the kind of pencil-grinding mathematics as has been represented here.

It bears noting that notwithstanding the claims to the contrary, no OCME analyst could accurately perform a LR of the type in Mr. Johnson's case by hand. It is too complex of a mixture. And that points up the prejudice, in that even the FST authors acknowledge that 3-person mixtures present opportunities for bad results, and stretch the limits of the experimental technology.[47] Those risks only increase with degraded samples, like those in Mr. Johnson's case. Tellingly, this is an area where FST's authors acknowledged in 2012, after jettisoning the degradation module and opting to utilize artificially low rates of drop-out and drop-in for all casework, regardless of the sample quality (hardly a conservative approach) that the program "might need improvement."[48]

VI. FST Does Not Reproduce Results as Expected

It is in the context of this history that recently it has come to light that, contrary to OCME's claims, FST has produced results that are unexpected and not reproducible. This evidence was discovered in the validation study. There, in two separate studies, one for a two-contributor mixture[49], one for a three-contributor mixture.[50] Given the data recorded in each study, FST should have produced uniform LR values. (i.e. $A; A$). Rather, what is observed is data that is incongruous (i.e. $B; C$).

---

[44] See Exhibit J Excerpt of Table 3A (page 1 of 22) and Exhibit K, excerpt of Table 3B (page 1 of 18) from OCME Validation Study.
[45] Mitchell 2 at 760.
[46] Ex. Summ., p. 11.
[47] See Mitchell 2 at 759 "[I]t would not be unusual to find alleles matching those of a non-contributor at many or all loci in a three-person sample."
[48] Id.
[49] Study 2C, Volume 19. See Exerpt of FST Validation Study 2C, Attached at Exhibit E.
[50],Study 3F, Volume 20. See Exerpt of FST Validation Study 3F, Attached at Exhibit F.

A. Study 2C.

In Study 2C, the OCME compared the profiles of two known contributors, D18 and D38, to two different mixed samples. It is clear from the documentation that in each mixed sample, both D18 and D38 produced a complete, 15 locus STR profile in that all alleles are present in the mixtures.[51] The evident variables, according to FST's available literature, indicate that the expected LR results should be identical for each profile in each mixed sample. Yet here the results are plainly incongruous.

In mixed sample D18_D38_2to1_500pg_NoV_2p_D_1sd_1000plus[52], FST produced a LR for D18 of 10.2 billion, while in D38_D18_2to1_500pg_NoV_2p_D_1sd_1000plus FST produced a LR of 11.7 billion. For D38, the difference in LR values was 18.1 million and 12.6 million, respectively.

B. Study 3F

In Study 3F, a three-person mixture, the discrepancies are more striking. There, several mixed samples were created from the profiles of known contributors D61, D62, and D63.[53] As with study 2C, all of the loci in each of the known contributors are present in each of the mixed samples. The published variables suggest that the LR values reported for each known contributor should match in each mixed sample.[54]

Again, the results varied, but on a much wider scale. While FST produced identical LRs for D63 of 126 million in each mixed sample[55], the program calculated LRs of 13,900 and 584 million for D62. Perhaps the most significant difference was calculated for D61, which FST calculated LR values of 271,000 and 90.7. Standing alone, this leap would take the LR within the OCME's own chosen subjective qualitative range down from "very strong" to "moderate" support for the putative prosecutor's hypothesis.

C. The Inadequacy of OCME's Reaction

These discrepancies are not minor. As with the hybridization of the profiles in the drop-out study, they were not reported to the Subcommittee, nor were they published in the literature. In actuality, these faults with FST appear to have been brought to light only in a recent state court case, State of New York v. Joseph Johnson, BX502/2014.

The OCME's response has been to flatly deny that the problem had anything to do with FST's source code. Parsing the program into "case" and "bulk calculation" functions, OCME Criminalist Craig O'Connor claimed that the problems with Studies 2C and 3F were

---

[51] See Exhibit E, Table 19C.2a.
[52] Id., Table 3A – 2 Person Mixtures.
[53] See Exhibit F, Table 20E.2.
[54] Because the template amounts differ by 100pg, the amounts could differ because of FST's "interpolation" function, although the reported effect of this variable is ambiguous in the literature. The subsequent claims of OCME Criminalist Craig O'Connor, however, acknowledge that the expected values should be identical. See infra p. 17.
[55] This consistent result would seem to strongly contra-indicate a data switch, or some other explanation that avoids harder questions going to the integrity of FST.

the equivalent of a typographical misprint.[56] Mr. O'Connor further claims to have "re-calculated the exact data specified in the studies. These re-calculations generated identical results, as expected, since these mixtures in fact contained what appear to be the same characteristics."[57]

D. Speculation and New Irreproducibility

Aside from the stridently personal tone throughout Mr. O'Connor's affidavit, two things are notable. First, however indirectly, Mr. O'Connor concedes the discrepancy identified in the studies.[58] The explanation he offers of a "misprint" or data error, however, suggests that the program somehow selected two out of three contributors in Study 3F to impose the observed results. This strains credulity.

Second, in performing recalculations on those studies, Mr. O'Connor has only increased, rather than decreased the uncertainty surrounding these studies and the integrity of FST. Looking at Study 3C. Prior to Mr. O'Connor's re-calculations, there were the expected results (A; A), and the observed results (B; C). Having run new calculations, Mr. O'Connor claims that the results are now "identical" and "as expected." Yet, he does not undertake to share what the new LR value he derived is.

Nevertheless, Mr. O'Connor can only mean he obtained values consistent with each other, that is "horizontally" across contributors (although plausibly also consistent with either (B) or (C). Thus, the new results could be (*B; B*), (*C; C*), a value originally expected (*A; A*), or some new value altogether (*D; D*).) However, because we cannot undo the discrepancy, Mr. O'Connor's new experiment only confirms a problem with "vertical" reproducibility. That is, at one time, for a reason, FST produced unexpected results. Today, according to Mr. O'Connor, based on the "exact data", FST produced different, putatively expected results.

This is hardly trivial. In the LR domain, without a ground truth, the only "expected" result that can be tracked is the concordance between the pairs of identical samples, a disruption of which would be represented in a horizontal discrepancy noted in Studies 2C and 3F. Disruptions of a vertical pair – that is, identical pairings run at different times – are no less troublesome as to the integrity of the code, but evidence of a vertical discrepancy in casework would be virtually impossible to track.

There certainly are other possible, obvious and more prosaic explanations available for the discrepancy: we learn from Mr. O'Connor that the OCME has conducted regular quality control runs on FST since its inception.[59] Have those runs identified source code

---

[56] See Exhibit H, Affidavit of Craig O'Connor, Submitted in People v. Joseph Johnson, Ind. 502-2014 (Sup. Ct. Bx. Cty. March 2, 2016) at ¶ 10. It is not precisely clear what Mr. O'Connor means by "identical" results. It can be assumed that the LR's are concordant with each other, "as expected." Whether the results obtained by Mr. O'Connor in 2015 are identical to any results from 2010 is left unanswered.

[57] Id.

[58] "[A]ny "inconsistency" in studies 2C and 3F can be equated to at most a typographical misprint and not an error in the program[.]" Affidavit of Craig O'Connor, filed in New York vs. Joseph Johnson, BX502/2014. Attached at Exhibit H.

[59] Exhibit H, at ¶ 8.

errors or bugs? How have those bugs affected casework? How have they implicated the validation studies? Given OCME's record of evasiveness regarding material issues and the source code we do not believe these questions can be answered accurately absent the relief requested.

Further, Mr. O'Connor's attempt to passing off the difficulty onto the validation study and away from casework is not satisfactory. As an initial matter, the claim that FST is actually two wholly separate and isolated computing elements[60] is entirely fresh, if inherently doubtful.[61] To construct a design as described would have doubled the programming burdens on FST's authors during its development and validation, a period during which the authors were willing to hybridize human STR profiles in the drop-out study with simulated loci, despite the obvious future limitations this time-saving step would place on their ability to extrapolate meaningful information from the data, let alone the reputational costs attendant to not disclosing this maneuver. The concept that the program is comprised of two separate programs, like Microsoft Word and Microsoft Excel, which share no common elements, only invites further inquiry.

Following on this, the claim that these discrepancies of such magnitude had no effect the false positive study is simply speculative. Notably, Mr. O'Connor offers no support for this statement. While it is certainly possible that the discrepancies were overlooked in the validation study, they could also have been willfully ignored. They were certainly not reported. But to extrapolate from there that thus no further errors existed or have taken place, that the software did function, has functioned, and continues to function without fault is to substitute, without inquiry, the best possible scenario as the sole explanation for all possible explanations. This is unacceptable.

The fact remains that the OCME's own validation study demonstrates that FST produces inconsistent LR values. No amount of stridency or revisionism can paper over that published account. OCME's claim that these problems could not affect casework eerily echo the circumstances encountered by Australian authorities, who because of what was characterized as a "minor miscode" in the probabilistic genotyping software tool STRmix, have been forced to restate the LR values in approximately 60 cases.[62] This evidence puts Mr. Johnson's right to inspect the true algorithm that produced the LR in his case into stark relief.

---

[60] Id. at ¶ 9.

[61] The sole apparent direct reference to the "bulk calculator", classically opaque on the assertion raised by Mr. O'Connor, appears in Mitchell 2 at p. 750: "This was achieved by running FST using each individual in a population database of 1246 non-contributors, collected by OCME and NIST as test profiles." (emphasis added).

[62] See http://www.couriermail.com.wu/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3flc59039039cfd1a2ef55af92b.(John Buckleton, the author of STRmix on the discrepancies with his program: "When we looked at the circumstances needed to cause this, we thought it was almost impossible. We can't replicate it[.]") Attached at Exhibit I, at p.2.

VII. Further Support in the Case Law

The child pornography cases cited by the Government regarding the disclosure of source code are either readily distinguishable or only further support our position requiring disclosure of FST source code. As an initial matter, child pornography is readily distinguishable from probabilistic genotyping software and the LR paradigm, as the existence of child pornography is objectively verifiable, traceable, and conclusively determinable through standard means. Conversely, LR statistics are theoretical numbers generated by highly sensitive, dependent equations that are ethereal, imaginary, and simply resist measurement by conventional means. .

Secondly, as the First Circuit pointed out in United States v. Chiaradio, 684 F.3d 265, 276 (1st Cir. 2012), the EP2P software at issue in that case is a closely-held "investigatory tool." Here, FST is quite literally an evidence-creating machine. Parsing the court's decision, two factors are operative in that case not present here. There was a trail that led directly to the defendant's computer. This trail was reproducible through reports, and the clear, objective existence of child pornography extant on the defendant's computer not only verified the accuracy and specificity of the EP2P process, but also inherently minimized the prejudice against him. Here, few of those factors are present. There is no such thing as an objective or objectively-verifiable LR statistic. Such numbers are self-referential.

Finally, unlike with E2P2, there is simply no lurking concern against disclosure that a sophisticated hacker might learn to frustrate investigative tools used to fight crime.

In United States v. Chiaradio, 684 F.3d 265 (1st Cir. 2012), the software traced IP addresses and digital fingerprints of files and it had no error rate (unlike FST) and had never yielded a false positive (unlike FST). Chiaradio, 684 F.3d at 278. In United States v. Cottom, 8:15CR239, 2015WL9308226 (D. Neb. Dec. 22, 2015), the Flash application was produced from a publicly available website and was "straightforward, simple." Cottom, 2015WL9308226 at *3. The defense in that case had access to the software itself, and was able to disassemble it into its functional parts. Then, the defense was able to observe and evaluate as it was executed to see if it did what it purported to do. Such an evaluation of FST would likely be sufficient for our purposes, but is not possible. First, the defense does not have FST. Second, even if we had it in hand, the fundamental difference between the objective nature of child pornography and variable likelihood ratios would likely frustrate the kind of "sandbox" review discussed in Chiardo. (Notably, in both cases cited in text by the Government, the District Court held a hearing regarding the source code.)

In another child pornography case, the Ninth Circuit came to the opposite conclusion from the First Circuit. United States v. Budziak, 697 F.3d 1105 (2012). The Court held that the District Court abused its discretion when it denied the defendant discovery of the same software as in Chiaradio. Id. at 1113. Like the Government has done in Mr. Johnson's case, the Government in Budziak argued that some discovery (in that case, computer logs) had been provided and the software being sought would not reveal anything else. Id. at 1112-13. The defense "presented arguments and evidences suggesting the materials disclosed by the FBI did not resolve all questions relevant to the defense" (in contrast to the defendant in Chiaradio). Id. at n.1. The Ninth Circuit held:

> In cases where the defendant has demonstrated materiality, the district court should not merely defer to government assertions that discovery would be fruitless. While we have no reasons to doubt the government's good faith in such matters, criminal defendants should not have to rely solely on the government's word that further discovery is unnecessary. This is especially so where, as here, a charge against the defendant is predicated largely on computer software functioning in the manner described by the government, and the government is the only party with access to that software.

Id. at 1112-13. See also United States v. Pirosko, 787 F.3d 358 (2015) (Sixth Circuit acknowledged the district court could have followed either Chiaradio or Budziak in deciding a motion to compel the same child pornography source code, depending on the defendant's presentation of error); United States v. Ocasio, 11CR2728, 2013WL2458617 (W.D. Tex. 2013) (granting a motion to compel source code sought from a third party pursuant to a Rule 17 subpoena in a child pornography case because the source code was directly relevant to the challenge brought by the defense (in that case, suppression)). As directed above, we have made a sufficient showing that the materials provided to date will not resolve all of the questions relevant to our Daubert motion.

The Government is correct when it states that "a defendant does not automatically need source code to challenge the reliability of expert testimony involving a computer program."[63] We have claimed no such thing. Nothing about this Rule 17 application is automatic. But FST is more than a mere computer program. Concealed within a black-box visible only to the government and its agents, FST assigns a weight of evidence – likely dispositive before any jury - to a complex and degraded mixed sample of DNA with such low quality that merely five years ago it would surely have resulted in Mr. Johnson's exclusion from the sample. The OCME has been ahead of this process at each step of the way. Its claims to prejudice should be unavailing upon this Court.

VIII. Conclusion

On this record, at best, FST has been qualitatively assessed with a level of inattention that allowed discordant results to escape notice during the validation study. The quantitative assessments, such as they are, cannot be assessed for completeness, but represent a miniscule proportion of the whole. This level of testing is hardly sufficient to effectively test the program for accuracy, and lends strength to observations of discordant results.

The premise that FST is reliable because the OCME claims it is reliable defies the logic of due process. The evidence of errors, which the OCME invites this court to accept as mere "typos" only begs the question of how the source code could be pristine. Every argument raised to obstruct Mr. Johnson's access, predicated on sound constitutional principle, has misread the facts of caselaw, exaggerated the rigor and substance of the testing to which FST was subjected, or has attempted to misdirect the courts attention from the true nature of this effort. We cannot interfere with FST's operation by perceiving its actual function; we have no commercial or other cognizable competitive interest against the OCME, and we have the obligation, right, and pressing need to inspect and test the reports and data by which the government's case is built and experts intend to base their testimony. Unless they intend to

[63] Letter from Government, May 19, 2016, p.2.

perform those calculations manually in this case, that basis is the true function of FST embedded in its code.

In short, Mr. Johnson seeks only to inspect and test the reliability of evidence offered against him. This is his right, and it should be granted.

For all of these reasons and those previously submitted, a subpoena for the source code should properly issue in accordance with Federal Rule of Criminal Procedure 17.

Respectfully submitted,

Christopher Flood
Sylvie Levine
Assistant Federal Defenders

cc: AUSA Jason Swergold
Kevin Johnson